BASEMENT

MASSACHUSETTS INSTITUTE of TECHNOLOGY
LIBRARIES

HD28
.M414
No. 1340-
82

Dewey

MASS. INST. Tc
JAN 11 1983
LIBRARIES

WORKING PAPER
ALFRED P. SLOAN SCHOOL OF MANAGEMENT

USING THE K-MEANS CLUSTERING METHOD AS A DENSITY ESTIMATION PROCEDURE

M. Anthony Wong
Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139

Working Paper #1340-82

MASSACHUSETTS
INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02139

USING THE K-MEANS CLUSTERING METHOD AS A DENSITY ESTIMATION PROCEDURE

M. Anthony Wong
Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139

Working Paper #1340-82

Key Words and Phrases: sample k-means clusters; histogram estimate; weak uniform consistency; computational requirement.

ABSTRACT

A random sample of size N is divided into k clusters that minimize the within cluster sum of squares locally. This k-means clustering method can be used as a quick procedure for constructing variable-cell historgrams that have no empty cell. A histogram estimate is proposed in this paper, and is shown to be uniformly consistent in probability.

1. INTRODUCTION

Let $X_1, X_2, \ldots, X_N$ be observations from some density f of a probability distribution F. In one dimension, the traditional method in estimating the univariate density f is the histogram. The asymptotic properties of the fixed cell historgram are given in Tapia and Thompson (1978). A major difficult of multivariate histograms, obtained by partitioning the sampled space into cells of equal size, is that there are too many cells with very few observations. Van Ryzin (1973) first proposed a variable cell histogram which is adaptive to

M.I.T. LIBRARIES
JAN 1 1 1983
RECEIVED

the underlying univariate density. Kim and Van Ryzin (1967) extended this method to the bivariate case; but the general procedure for the multivariate case is very complicated.

On the other hand, the theoretically sound density estimation techniques like the kernel method (Parzen, 1962) and the kth nearest neighbor method (Loftsgaarden and Quensenberry, 1965) have computational problems when applied to large data sets. (For the asymptotic consistency of these techniques, see for example, Devroye and Wagner (1977), Moore and Yackel (1977), and Silverman (1978).) Although the statistical justification of these density estimates require N very large, the number of computations is usually $O(N^2)$ which begins to be onerous for N over 500. In this paper, it is proposed that the widely-used k-means clustering technique can be regarded as a practical and convenient way of obtaining variable cell histograms in one or more dimensions; the computational requirement of this algorithm is $O(Nk)$, where k is the number of cells or clusters.

Suppose that the observations $x_1, \ldots, x_N$ are partitioned into k groups with means $u_1, u_2, \ldots, u_k$ such that no movement of an observation from one group to another will reduce the within groups sum of squares

$$WSS(N) = \sum_{i=1}^{N} \min_{1 \le j \le k} || x_i - u_j ||^2.$$

This technique for division of a sample into k clusters to minimize the within group sum of squares locally is known in the clustering literature as k-means. In one dimension, the partition will be specified by $k-1$ cutpoints; the observations lying between common cutpoints are in the same group. See Hartigan (1975) for a detailed description of the k-means technique, and see Hartigan and Wong (1979) for an efficient computational algorithm. The asymptotic properties of k-means as a clustering procedure (as N approaches ∞ with k fixed) have been studied by

0745074

MacQueen (1967), Hartigan (1978), and Pollard (1981). In this paper, it is shown that the k-means procedure can be used to construct a histogram estimate of the underlying density function.

In Section 2, using the asymptotic properties of k-means clusters (when $k \to \infty$ with N) given in Wong (1980), it is shown that the proposed histogram estimate is uniformly consistent in probability in one dimension. The multivariate case requires further investigation as the generalization of the univariate consistency result to many dimensions is not straightforward. However, empirical examples are given in Section 3 to illustrate the potential of k-means as a practical density estimation procedure for large multivarate data sets. Some conclusing remarks are also given in Section 3.

2. WEAK UNIFORM CONSISTENCY OF THE K-MEANS HISTOGRAM ESTIMATE

In this section, a k-means histogram estimate of an unknown univariate density function is proposed which is shown to be uniformly consistent in probability.

Let $x_1, \ldots, x_N$ be observations from a density function f which is positive and has four bounded derivatives in $[a,b]$. Suppose that the N observations are grouped into k_N clusters with means $u_1 < u_2 < \ldots < u_{k_N}$ and within-cluster sums of squares $WSS_1, \ldots, WSS_{k_N}$ such that the within groups sum of squares

$$WSS(N) = \sum_{j=1}^{k_N} WSS_j(N) = \sum_{i=1}^{N} \min_{1 \le j \le k_N} ||x_i - u_j||^2$$

of this locally optimal k_N-partition cannot be decreased by moving any single observation from its present cluster to any other cluster. Let $I_j = [y_{j-1}, y_j]$ be the set of points in $[a,b]$ closer to u_j than to any other cluster mean. Then $\{I_1, \ldots, I_{k_N}\}$ is the k_N-partition of $[a,b]$ defined by the cluster means $u_1, \ldots, u_{k_N}$, and $a = y_0 < y_1 < \ldots < y_{k_N - 1} < y_{k_N} = b$ are the cutpoints of this partition. Denote the size of the jth cluster

interval of this partition by e_j and let the number of observations in the jth cluster by n_j. Then it is shown in Wong (1980) that if $k_N = o([N/\log N]^{1/3})$,

$$\max_{1\le j\le k_N} |e_j\, k_N\, f_j^{1/3} - \int_a^b f(x)^{1/3}\, dx| = o_p(1), \tag{2.1}$$

where f_j is the density at the midpoint of the jth cluster interval,

$$\max_{1\le j\le k_N} |n_j\, N^{-1} k_N\, f_j^{-2/3} - (\int_a^b f(x)^{1/3}\, dx)| = o_p(1) \tag{2.2}$$

and

$$\max_{1\le j\le k_N} |\, 12\, WSS_j\, N^{-1}\, k_N^3 - (\int_a^b f(x)^{1/3} dx)^3| = o_p(1). \tag{2.3}$$

From (2.1), (2.2), and (2.3) respectively, by putting $G = \int_a^b f(x)^{1/3}\, dx$, we have uniformly in $1\le j\le k_N$,

$$e_j = G\, k_N^{-1}\, f_j^{-1/3}\, [1 + o_p(1)] \tag{2.4}$$

$$n_j = G\, N\, k_N^{-1}\, f_j^{2/3}\, [1 + o_p(1)] \tag{2.5}$$

and $$WSS_j = \frac{1}{12}\, G^3\, N\, k_N^{-3}\, [1 + o_p(1)] \tag{2.6}$$

Therefore, in constructing a histogram to estimate an unknown density function f which vanishes outside the finite interval $[a,b]$, equation (2.4) indicates that the k-means procedure would partition $[a,b]$ in such a way that the sizes of the intervals are adaptive to the underlying density; the intervals are large where the density is low while the intervals are small where the density is high. It follows that the k-means procedure can be regarded as a useful tool for constructing variable-cell histograms.

Define the density estimate at a point x by

$$f_N(x) = n_j^{2/3}/N(12\ WSS_j)^{1/2},\ y_{j-1} \le x < y_j \text{ for } 1 \le j \le k_N, \qquad (2.7)$$

Then from (2.5) and (2.6), we have

$$f_N(x) = (GN\ k_N^{-1}\ f_j^{2/3})^{3/2}\ [1 + o_p(1)]/N(G^3 N k_N^{-3})^{1/2}\ [1 + o_p(1)]$$

$$= f_j\ [1 + o_p(1)], \text{ uniformly in } 1 \le j \le k_N.$$

Since f is uniformly continuous,

$$\sup_{a \le x \le b} \left| f_N(x) - f(x) \right| = o_p(1).$$

And we have shown that the histogram estimate f_N is uniformly consistent in probability.

3. EMPIRICAL ANALYSIS OF THE K-MEANS HISTOGRAM ESTIMATE

The results in Section 2 indicate that the k-means procedure can be used to construct a uniformly consistent histogram estimate of an unknown univariate density f provided that k_N is chosen such that k_N is of order $o([N/\log N]^{1/3})$. An empirical study was performed to examine the performance of the k-means histogram estimate (see Wong 1979), in which the effectiveness of various choices of k_N was also compared. There is empirical evidence that the choice $k_N = 4N^{0.3}$ is effective over a range of sample sizes for various normal mixture densities. Two examples are given in Figure 1 and Figure 2, in which the performance of the k-means estimate (k=40) is illustrated by using 1000 generated observations from two different normal mixtures. The CPU time consumed on the IBM 370/158 for the two examples are 12.95 seconds and 14.18 seconds respectively.

A major difficulty of the usual histogram is that when multivariate histograms are constructed by partitioning the sampled space into cells of equal size, there are too many empty cells. One desirable feature of the k-means procedure is that it provides

a practical and convenient way of obtaining a k-partition of the multivariate sample, or equivalently, the multidimensional sampled space. Consequently, histogram estimates of the density over these k cells or regions (whose sizes are conjectured to be adaptive to the underlying density) can be obtained. However, the uniform consistency of such a multivariate histogram has not been established, and much work has yet to be done to investigate the asymptotic properties of k-means partitions of samples from multidimensional distributions.

Empirical bivariate examples are included here to illustrate the potential of the k-means technique as a practical density estimation procedure for large multivariate data sets. The three generated data sets are samples of size 1000 from (1) a bivariate normal with mean (0,0) and covariance matrix $\binom{10}{01}$; i.e., BVN [(0,0), $\binom{10}{01}$], (2) the mixture $\frac{1}{2}$BVN [(0,0), $\binom{10}{01}$] + $\frac{1}{2}$BVN [(3,3), $\binom{10}{01}$], and (3) the mixture $\frac{1}{2}$BVN [(0,0), $\binom{90}{04}$] + $\frac{1}{2}$BVN [(0,6), $\binom{90}{04}$]. The density estimates ($f_N(x) \propto n_i^{1+p/2} WSS_i^{-p/2}$; p=2 for bivariate data) over the k=40 clusters obtained by k-means are given respectively in Figures 3, 4, and 5. The results suggest that k-means is a useful tool for estimating density. Moreover, the computational requirement is only $O(Nk)$, which is considerably less prohibitive than the usual kernel and nearest neighbor techniques which require $O(N^2)$ computations; the average CPU time on the IBM 370/158 for the three bivariate examples is 22.32 seconds.

ACKNOWLEDGEMENTS

This research was supported in part by the National Science Foundation, Grant No. NCS75-08374. The author wants to thank John A. Hartigan for many useful discussions.

BIBLIOGRAPHY

Blashfield, R.K., and Aldenderfer, M.S. (1978). The Literature on Cluster Analysis. Multivariate Behavioral Research, 13, 271-295.

Devroye, L.P., and Wagner, T.J. (1977). The strong uniform consistency of nearest neighbor density estimates. Annals of Statistics, 5, 536-540.

Hartigan, J.A. (1975). Clustering Algorithms. New York: John Wiley and Sons.

_____(1978). Asymptotic distributions for clustering criteria. Annals of Statistics, 6, 117-131.

_____, and Wong, M.A. (1979). Algorithm AS136: A K-means clustering algorithm. Applied Statistics, 28, 100-108.

Kim, B.K., and Van Ryzin, J. (1976). A Bivariate Histogram Estimator, Technical Report No. 444, Dept of Statistics, University of Wisconsin, Madison.

Loftsgaarden, D.O., and Quensenberry, C.P. (1965). A nonparametric estimate of a multivariate density function. Annals of Mathematical Statistics, 35, 1049-1051.

MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Probability and Statistics, 281-297.

Moore, D.S. and Yackel, J.W. (1977). Consistency properties of nearest neighbor density function estimators. Annals of Statistics, 5, 143-154.

Parzen, E. (1962). On estimation of a probability density function and mode. Annals of Mathematical Statistics, 33, 1065-1076.

Pollard, D. (1981). Strong consistency of k-means clustering. Annals of Statistics, 9, 135-140.

Silverman, B.W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. Annals of Statistics, 6, 177-184.

Tapia, R.A., and Thompson, J.R. (1978). Nonparametric Probability Density Estimation. Baltimore: The John Hoptins University Press.

Van Ryzin, J. (1973). A histogram method of density estimation. Communications in Statistics, 2, 493-506.

Wong, M.A. (1979). Hybrid Clustering. Unpublished Ph.D. thesis Department of Statistics, Yale University.

FIG 1: Density Estimates obtained by k-means using k=40.

(1000 observations from ½ N (0,1) + ½ N (3,1).)

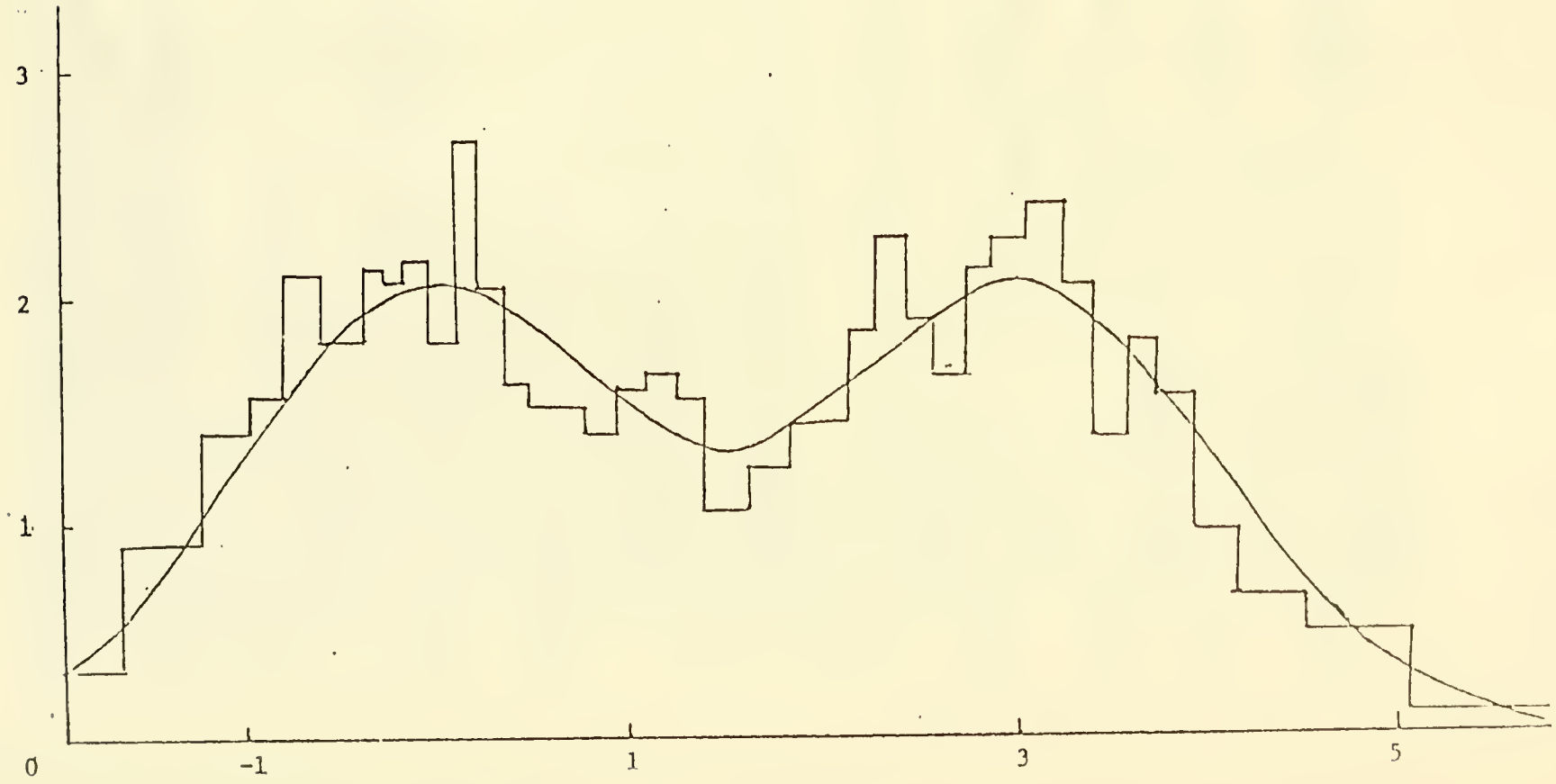


FIG 2: Density Estimates obtained by k-means using k=40.

(1000 observations from ½ N (0,1) + ½N (4,1).)

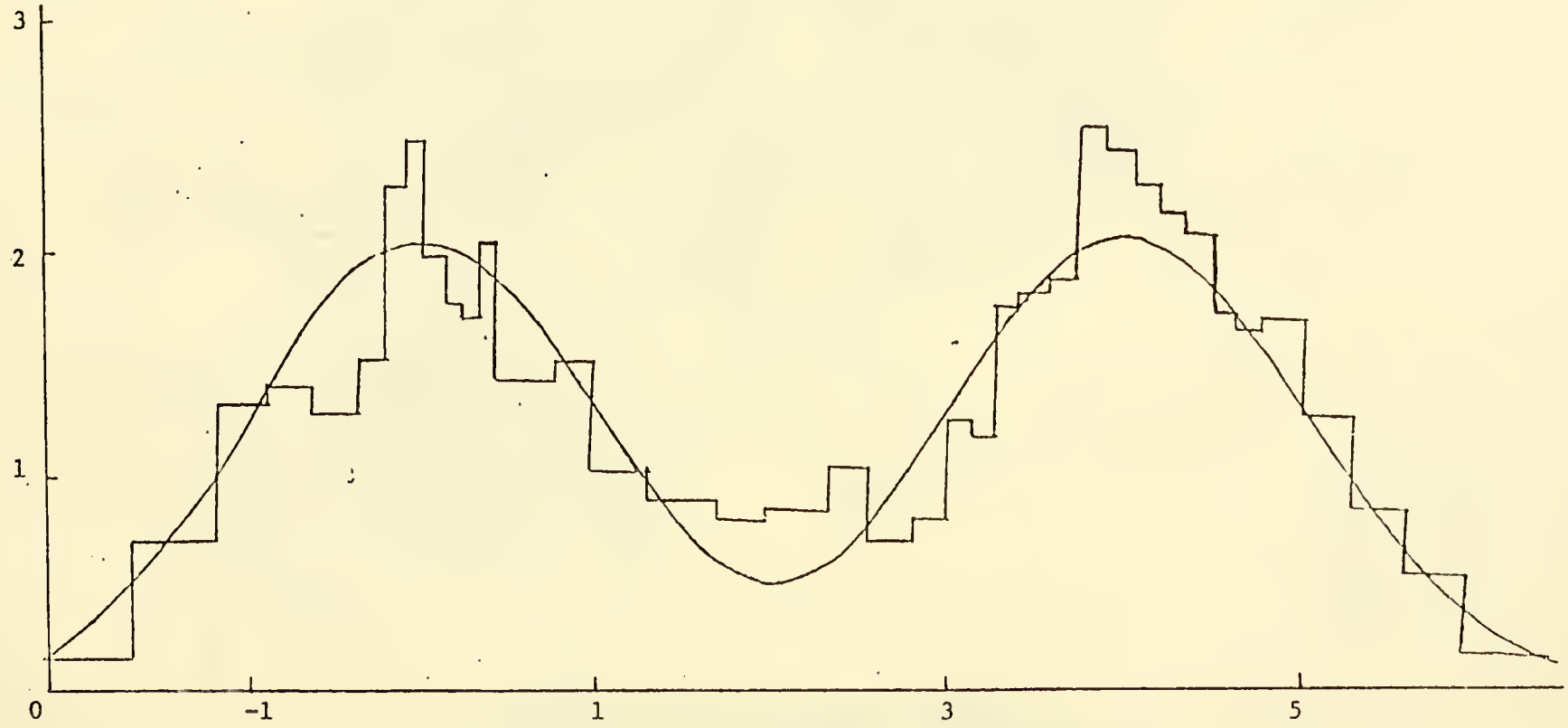


FIG 3: Density Estimates obtained by k-means using k=40.
(1000 observations from BVN[(0,0),$\begin{pmatrix}1&0\\0&1\end{pmatrix}$].)

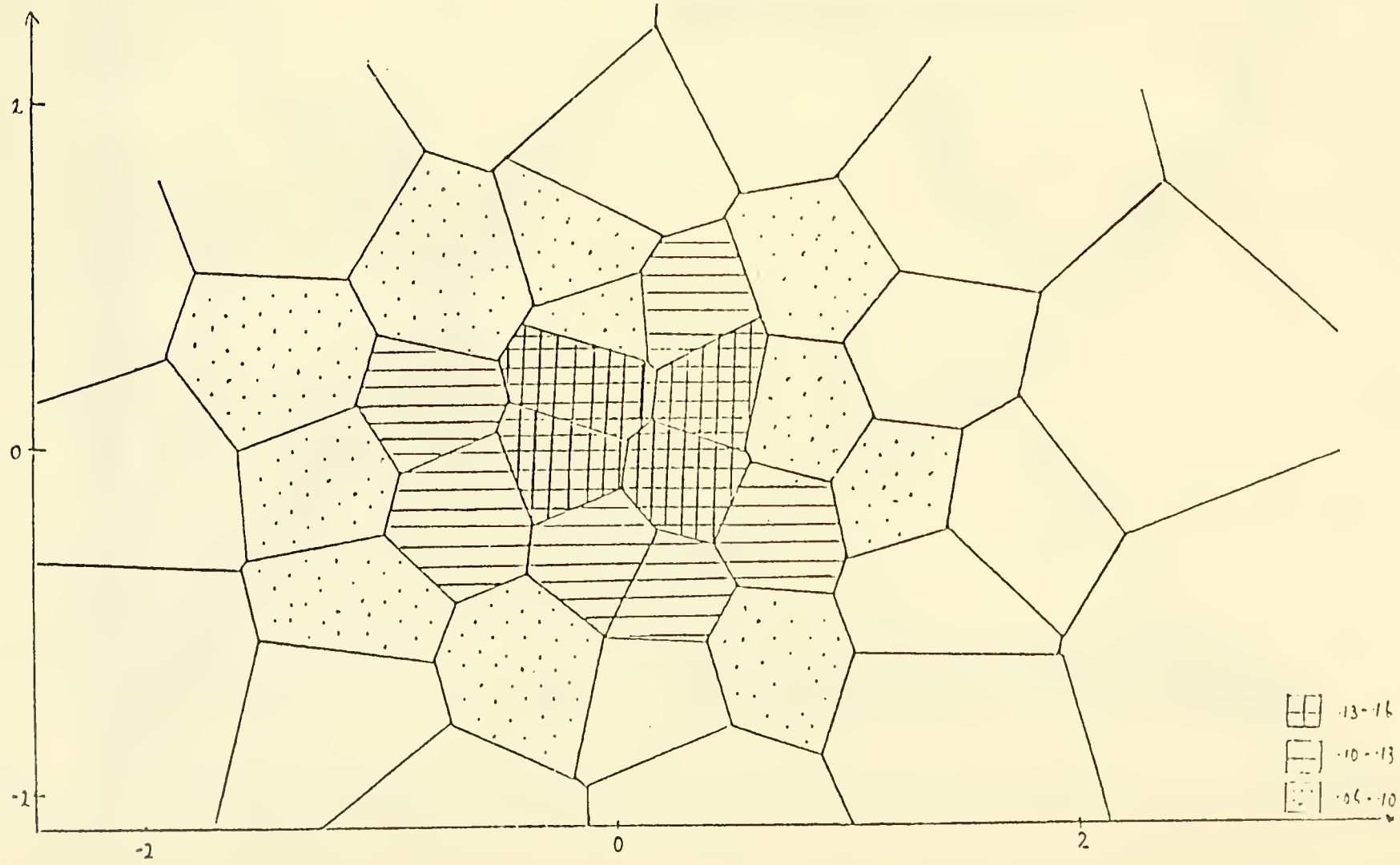


FIG 4: Density Estimates obtained by k-means using k=40.
(1000 observations from 1/2 BVN $[(0,0), \binom{10}{01}]$ + 1/2 BVN $[(3,3), \binom{10}{01}]$.)

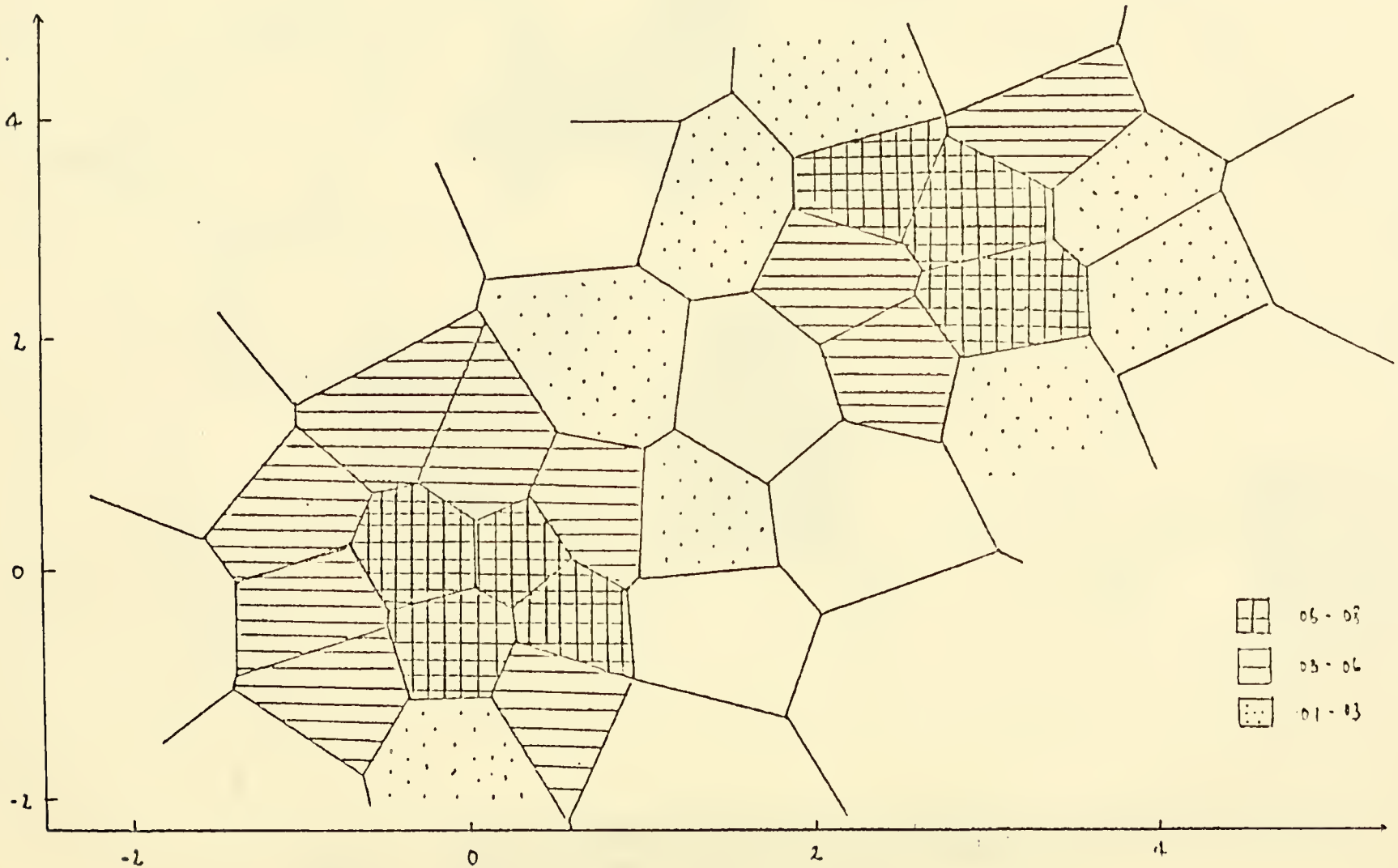
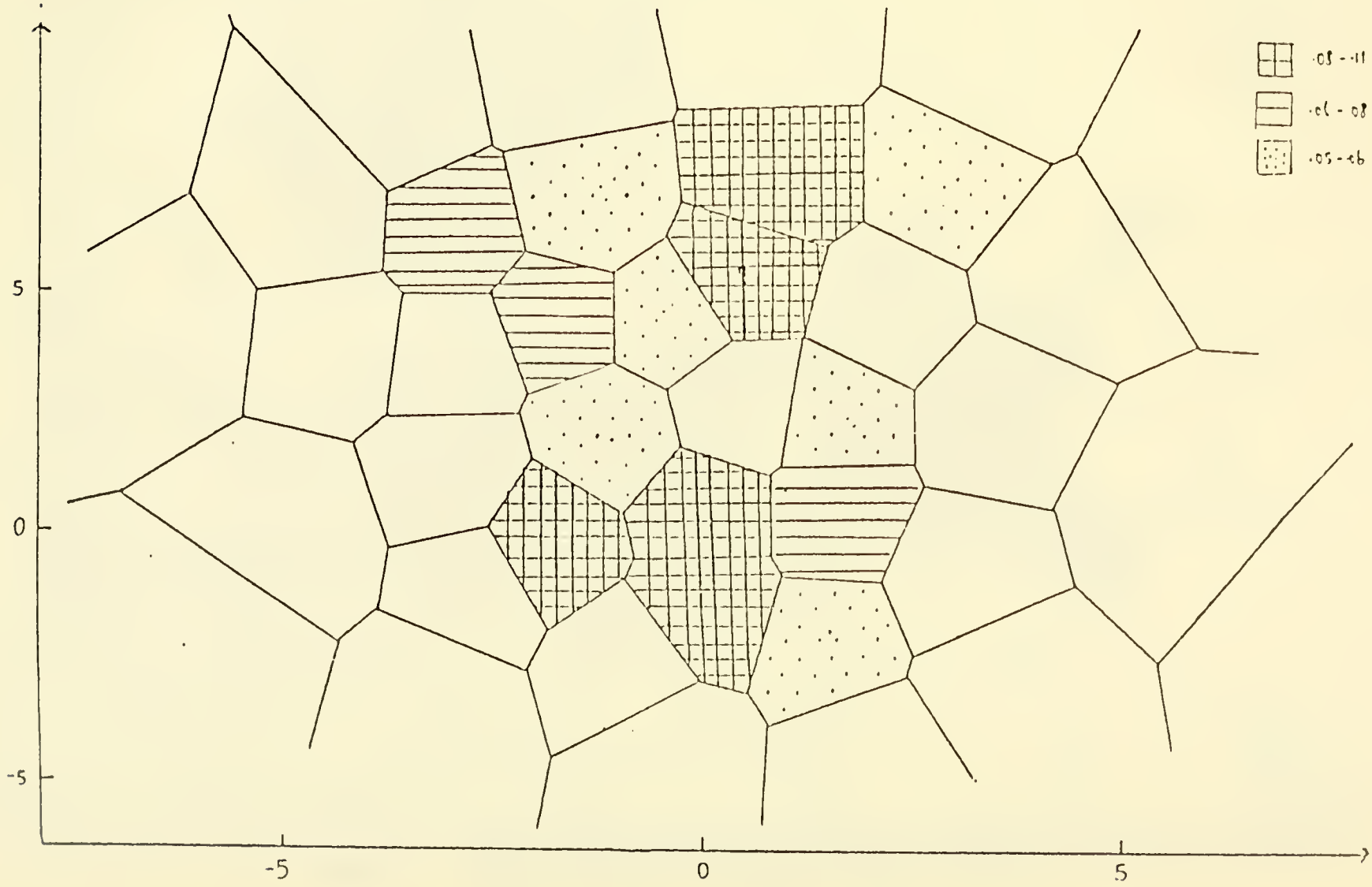


FIG 5: Density estimates obtained by k-means using k=40.
(1000 observations from 1/2 BVN $[(0,0), \binom{90}{04}]$ + 1/2 BVN $[(0,6), \binom{90}{04}]$.)



BASEMENT
Date Due

Lib-26-67

HD28.M414 no.1340- 82
Wong, M. Antho/Using the K-means clust
745074 D*BKS 00136957
3 9080 002 047 428